Applications of Minimum Bayes-Risk Decoding to LVCSR

Vaibhava Goel and William Byrne

Center for Language and Speech Processing
Johns Hopkins University, Baltimore, MD, U.S.A.

Loss Functions

- **Quantitative measure of goodness of recognizer output**

- **Task dependent**

							Loss (Truth, Hyp)	
							Words	Names
TRUTH	:	MARK	MUST	GO	THERE	⋯		
HYP I	:	MARK	BEST	GO	THERE	⋯	1	0
HYP II	:	DARK	MUST	GO	THERE	⋯	1	1

- **Examples**
  - **Conventional: Word Error Rate**
  - **Named Entity Extraction: F-measure, Slot Error Rate**
  - **Keyword Spotting, Information Retrieval: Keyword Error Rate**
  - **Dialogue Systems: Quality of dialogue**

What is a Good Recognizer $\delta(A)$?

- **When used in real scenarios, it minimizes Bayes risk**

$$E_{P(W,A)}\left[l(W,\delta(A))\right]$$

- **Minimum Bayes-risk (MBR) recognizer on an utterance A**

$$\delta(A) = \underset{W' \in \mathcal{W}_h}{\operatorname{argmin}} \sum_{W \in \mathcal{W}_e} l(W,W')P(W|A)$$

DEFINITIONS

- $\delta(A)$ **: Decision for utterance A.**
- $\mathcal{W}_h$ **: Hypothesis space**
  - **Decoder's choices are limited to this space**
- $\mathcal{W}_e$ **: Evidence space**
  - **Set of possible generators of the acoustic data**
  - **Derived using the acoustics and the models.**
- $P(W|A)$ **: Evidence distribution**
  - **Probability of elements of evidence space given the acoustics A.**

Approximate Implementations

- **Hypothesis and evidence spaces could be N-best lists or lattices.**

- **Hypothesis space could even be picked arbitrarily.**

- **N-best list rescoring** [1]

$$\delta(A) = \underset{W' \in \mathcal{W}_{nb1}}{\mathrm{argmin}} \sum_{W \in \mathcal{W}_{nb2}} l(W, W') P(W|A)$$

- **Lattice rescoring** [2]

$$\delta(A) = \underset{W' \in \mathcal{W}_{lat}}{\mathrm{argmin}} \sum_{W \in \mathcal{W}_{lat}} l(W, W') P(W|A)$$

  – **Lattice rescoring accomplished with a prefix tree based multi-stack A^* search.**

[1] A. Stolcke, Y. Konig, and M. Weintraub, “Explicit word error minimization in N-best list rescoring,” Eurospeech-97, pp. 163–165, Rhodes, Greece, 1997.

[2] V. Goel and W. Byrne, “Minimum Bayes-risk automatic speech recognition,” To appear in Computer Speech and Language, 2000.

Transcription of Speech & Keyword Spotting

- **Transcription :** $l(W, W')$ **= Weighted Levenshtein distance (L).**
- **Keyword spotting :** $l(W, W')$ **= L between keywords of** W **and** W'**.**

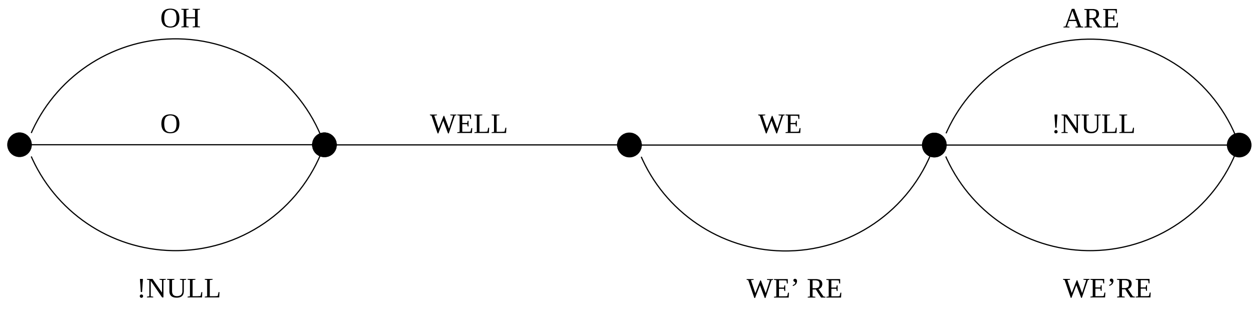
- **Test-set : JHU-WS97 dev-test set. 38 conv. sides, 2427 utterances.**
- **Thanks to Dimitra Vergyri for lattices.**
- **Results:**

Loss function	Decoding strategy	SER	WER	KER
SER	MAP	65.9	38.5	43.2
WER	N-best	66.8	37.9	43.0
	A-star	66.8	37.5	42.4
KER	N-best	N/A	N/A	42.0
	A-star	N/A	N/A	41.4

ROVER : Summary [3]

- **Voting on outputs of N_s recognizers**
- **Iteratively construct a word transition network (WTN) by adding one system at a time**

OH (0.3)	WELL (0.9)	WE (0.7)	ARE (0.6)
O (0.2)	WELL (0.7)	WE’RE (0.7)	!NULL (0.7)
!NULL (0.7)	WELL (1.0)	WE (0.8)	WE’RE (0.6)



- **From correspondence set j, select the word with maximum $S(w,j)$ defined as**

$$
\begin{aligned}
\textbf{Nist1 :}\ S(w,j) &= N(w,j)/N_s \\
\textbf{Nist2 :}\ S(w,j) &= (\alpha N(w,j) + (1-\alpha)C(w,j))/N_s \\
\textbf{Nist3 :}\ S(w,j) &= \alpha(N(w,j)/N_s) + (1-\alpha)\max_{w \in CS_j} C(w,j)
\end{aligned}
$$

[3] J. Fiscus, “A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER),” IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 347–354, Santa Barbara, CA, 1997.

ROVER is an MBR Decoder

- **Loss function :**
  - **The loss function l_{ROVER} is obtained by summing the word errors over correspondence sets.**
  - **l_{ROVER} approximates the word error rate.**
  - **The alignment to compute WER is specified by the WTN. For example :**

OH (0.3)	WELL (0.9)	WE (0.7)	ARE (0.6)
O (0.2)	WELL (0.7)	WE'RE (0.7)	!NULL (0.7)
!NULL (0.7)	WELL (1.0)	WE (0.8)	WE'RE (0.6)

- **Hypothesis space :**
  - **Set of all possible paths through the WTN.**
- **Evidence space and distribution :**
  - **Evidence space is the union of evidence spaces of N_s systems.**
  - **Let $P_k(W|A)$ be the evidence distribution over the evidence space of the k^{th} system.**
  - **Assume that the confidence $C_k(W^j, j)$ are the posterior probability of W^j under $P_k(W|A)$.**

Evidence Distribution for ROVER

- **The following evidence distribution underlies the voting procedures of Nist1 and Nist2**

$$P(W|A) = \frac{1}{N_s} \sum_{k=1}^{N_s} \left[\alpha \hat{P}_k(W|A) + (1-\alpha) P_k(W|A) \right].$$

where

$$\hat{P}_k(W|A) = \begin{cases} 1 \textbf{ if } W = \delta_k(A), \\ 0 \textbf{ otherwise} \end{cases}$$

- $S(w,j)$ **for Nist1 and Nist2**

$$\begin{aligned} \textbf{Nist1}: S(w,j) &= N(w,j)/N_s \\ \textbf{Nist2}: S(w,j) &= (\alpha N(w,j) + (1-\alpha) C(w,j))/N_s \end{aligned}$$

- **Thus, ROVER is an MBR procedure under the above specified loss function (approximate WER), hypothesis space, evidence space, and evidence distribution. This may be why it is effective at reducing the word error rate!**

- **We have recently formulated a *segmental MBR decoding* scheme [4]. ROVER and lattice based voting scheme of Mangu et.al [5] can be shown to be instances of segmental MBR decoding.**

[4] V. Goel and W. Byrne, "Recognizer output voting and DMC in minimum Bayes-risk framework," Research Notes No. 40, CLSP, JHU, 2000.

[5] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus among words: lattice-based word error minimization," Eurospeech-99, pp. 495–498, Budapest, Hungary, 1999.

Towards e-ROVER : Enlarged Hypothesis Space

Limitations of ROVER, as originally formulated

- **Hypothesis space is limited to WTN constructed from 1-best of each system.**
- **Loss function l_{ROVER} over-estimates word error rate.**

Step 1: Keep everything else as in ROVER, enlarge hypothesis space.

- **Allow a bigger hypothesis space by constructing a WTN from N-best lists of N_s recognizers (N-best ROVER).**
- **Due to a larger hypothesis space, N-best ROVER has a smaller expected loss than ROVER under any loss function.**
- **Loss function l_{ROVER} approximates WER, therefore we expect N-best ROVER to yield lower WER than ROVER!**

Step 2 : Improved WER Approximation

- l_{ROVER} is restricted to loss based on alignment derived from WTN.
- Improve this approximation by allowing correspondence sets to have strings of more than one word.
- Example
  - WTN constructed from 3 recognizer outputs

OH	WELL	WE	ARE
O	WELL	WE’RE	!NULL
!NULL	WELL	WE	WE’RE

  - Allowing multiple words in correspondence sets

OH	WELL	WE ARE
O	WELL	WE’RE
!NULL	WELL	WE WE’RE

e-ROVER

- Construct a WTN using N-best lists from N_s systems.
- Decide on correspondence sets to 'pinch' on and thus identify correspondence sets with word strings in them.
  - Example

OH	WELL	WE	ARE	HERE	AFTER	ALL
O	WELL	WE'RE	!NULL	HERE	AFTERALL	!NULL
!NULL	WELL	WE	WE'RE	HERE	AFTER	ALL

- Pick a hypothesis from each correspondence set according to a segmental minimum Bayes-risk procedure.
  - The hypothesis space, evidence space, and evidence distribution stay that of ROVER.
  - WER <= Loss under e-ROVER <= l_{ROVER}.
- Expected WER under e-ROVER <= that under N-best ROVER !!

Preliminary Results

- **Multi-lingual language independent acoustic modeling for Czech [6].**
- **Joint work with Shankar Kumar.**
- **3 Systems:**
  - **Czech triphone acoustic models.**
  - **Czech N-best lists rescored with English triphone acoustic models adapted to Czech.**
  - **English triphone acoustic models adapted to Czech.**
- **250-best from each system.**
- **Baseline WER : 29.42, 29.22, and 35.24.**
- **ROVER : 26.68 (-2.54)**
- **N-Best ROVER : 25.95 (-3.27)**
- **e-ROVER : 25.43 (-3.79)**

[6] W. Byrne et.al., “Towards language independent acoustic modeling,” To appear in ICASSP00, Istanbul, Turkey, 2000.